WRDC-TR-89-7006
Volume II

ASSESSMENT OF CREW WORKLOAD MEASUREMENT METHODS, TECHNIQUES AND PROCEDURES

Volume II - Guidelines for the Use of Workload Assessment Techniques in Aircraft Certification

William H. Corwin, Diane L. Sandry-Garza
Michael H. Biferno, George P. Boucek, Jr.

DOUGLAS AIRCRAFT COMPANY
3855 LAKEWOOD BLVD
LONG BEACH, CALIFORNIA 90846-0001

BOEING COMMERCIAL AIRPLANES
P. O. BOX 3707
SEATTLE, WASHINGTON 98124-2207

September 1989

Final Report for Period July 1986 - February 1989

COCKPIT INTEGRATION DIRECTORATE
WRIGHT RESEARCH AND DEVELOPMENT CENTER
AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO 45433-6553

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely related Government procurement operation, the United States Government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture use, or sell any patented invention that may in any way be related thereto.

This report has been reviewed by the Office of Public Affairs (ASD/PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

Michael H. Pharaoh
Wing Commander, Royal Air Force

PAUL E. BLATT
Technical Director

FOR THE COMMANDER

EUGENE A. SMITH, Col, USAF
Director
Cockpit Integration Directorate

If your address has changed, if you wish to be removed from our mailing list, or if the addressee is no longer employed by your organization please notify WRDC/FIGX *, W-PAFB, OH 45433 to help us maintain a current mailing list.*

Copies of this report should not be returned unless return is required by security considerations, contractual obligations, or notice on a specific document.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS NONE
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) WRDC-TR-89-7006 VOLUME II
6a. NAME OF PERFORMING ORGANIZATION DOUGLAS AIRCRAFT COMPANY 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION WRIGHT RESEARCH DEVELOPMENT CENTER COCKPIT INTEGRATION DIRECTORATE WRDC/KT
6c. ADDRESS (City, State, and ZIP Code) LONG BEACH CA 90846	7b. ADDRESS (City, State, and ZIP Code) WRIGHT-PATTERSON AFB OH 45433-6553
8a. NAME OF FUNDING/SPONSORING ORGANIZATION FEDERAL AVIATION ADMINISTRATION 8b. OFFICE SYMBOL (If applicable) APM 430	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F33615-86-C-3600
8c. ADDRESS (City, State, and ZIP Code) 800 INDEPENDENCE AVENUE WASHINGTON DC 20591	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 62201F; PROJECT NO. 2403; TASK NO 04; WORK UNIT ACCESSION NO. 50
11. TITLE (Include Security Classification) Assessment of crew workload measurement methods, techniques and procedures. Vol II - Guidelines for the Use of Workload Assessment Techniques in Aircraft Certification	
12. PERSONAL AUTHOR(S) William H. Corwin, Diane L. Sandry-Garza, Michael A. Biferno, George P. Boucek	
13a. TYPE OF REPORT Final 13b. TIME COVERED FROM Jul 86 TO Feb 89	14. DATE OF REPORT (Year, Month, Day) 89, Sep 12 15. PAGE COUNT 51
16. SUPPLEMENTARY NOTATION Supported in part by the FAA	

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP
05	09	

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Subjective Measures
Performance Measures
Physiological Measures

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The Final Report, Volumes One and Two, summarizes the work completed under the FAA/USAF workload contract F33615-86-C-3600. The purpose of Volume Two is to present specific guidelines and recommendations for evaluating workload certification plans. No attempt is being made to provide alist of simple-to-follow directions for the generation of an aircraft workload certification plan, as this is the responsibility of the manufacturer. Volume One summarizes the activities leading up to and including two user community workshops and two simulation studies conducted at the Man-Vehicle Systems Research Facility, NASA-Ames Research Center.

The workload assessment techniques are discussed by domain area: Submective, Physiological, Performance, and Analytic techniques. The distinction by domain is convenient because of the methods and equipment in common amont techniques within a domain.

Evaluation criteria for assessing a workload certification plan includes treatment of the validity, reliability, and applicability of candidate workload measures. For a workload

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a NAME OF RESPONSIBLE INDIVIDUAL Michael H. Pharaoh	22b. TELEPHONE (Include Area Code) (513) 255-8279 22c. OFFICE SYMBOL WRDC/KT

DD Form 1473, JUN 86 *Previous editions are obsolete.* SECURITY CLASSIFICATION OF THIS PAGE

Unclassified

measure to demonstrate validity, it must be able to discriminate among varying task demands imposed upon the flightcrew. In order for a measure to demonstrate reliability, it should provide the same results with repeated applications. Applicability is simply the ability for workload to be assessed in an aircraft flightdeck environment.

Volume Two provides guidelines, criteria, and candidate workload measures for aircraft certification based upon empirical findings from the simulation studies conducted at NASA-Ames. workload measures demonstrating validity and reliability include: Subjective techniques (SWAT, NASA-TLX, and the Bedford scale), Heart Rate, and Control Input Activity for the wheel (ailerons) and column (elevator). These workload measures should not be considered an exhaustive list of workload assessment techniques. Other workload measures exist which we were unable to evaluate because of budget and time limitations and they may be just as valid and reliable as the ones listed above. Rather than just presenting an exhaustive list of workload measures for aircraft certification, this volume is intended to provide a methodology by which workload measures can be evaluated for validity and reliability. In a few years, many of the current state-of-the-art workload measures may become obsolete. The contents of Volume Two allow for the evaluation of current, and yet to be developed, workload assessment techniques.

Advantages and liabilities of the techniques employed in the simulation studies (reported in Volume One) are discussed in implementation sections. Some of the previous work reported by others is noted. Finally, the process of evaluating the workload assessment flights includes sections addressing scenario description, scenario evaluation criteria, and relation of Workload to FAR 25.1523, Appendix D, requirements.

The emphasis of this document is to provide guidelines for those involved with determining the adequacy of a certification plan for flightdeck workload, most notably Aircraft Certification Officers of the FAA. The information presented in these volumes to support the FAA users and provide clear guidance to workload specialists on aircraft programs.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

GLOSSARY

ANOVA	Analysis of Variance
ATC	Air Traffic Control
ATIS	Automated Terminal Information System
ATP	Airline Transport Pilot
BCA	Boeing Commercial Airplanes
DAC	Douglas Aircraft Company
EB	Eyeblink Rate
FAA	Federal Aviation Administration
FAR	Federal Aviation Regulation
FM	Frequency Modulation
HR	Heart Rate
HRV	Heart Rate Variability
HSD	Tukey's Honestly Significant Difference
Hz	Hertz
ILS	Instrument Landing System
IBI	Interbeat Interval
IMC	Instrument Meteorological Conditions
LED	Light Emitting Diode
MEL	Minimum Equipment List
MSe	Mean Square Error
MVSRF	Man-Vehicle System Research Facility
NASA	National Aeronautics and Space Administration
NOTAMS	Notice to Airmen
OAK	Oakland International Airport
OWS	Overall Workload Score
PCA	Principal Component Analysis

GLOSSARY
(Continued)

PF	Pilot Flying
PNF	Pilot Not Flying
PSE	Pilot Subjective Evaluation
PTT	Push to Talk
SBP	Power Spectral Analysis (Blood Pressure Component)
SCK	Stockton Municipal Airport
SD	Standard Deviation
SELCAL	Selective Calling
SFO	San Francisco International Airport
SID	Standard Instrument Departure
SIGMET	Significant Meteorological Conditions
SMF	Sacramento Municipal Airport
SRS	Power Spectral Analysis (Respiration Component)
STAR	Standard Terminal Arrival
STPA	Secondary Task Probe Accuracy
STRT	Secondary Task Reaction Time
SWAT	Subjective Workload Assessment Technique
TLA	Timeline Analysis
TLX	NASA-Time Load Index
TOC	Top of Climb
TOD	Top of Descent
USAF	United States Air Force
VMC	Visual Meteorological Conditions

PREFACE

This report is the result of two years of research sponsored by the USAF and the FAA directed toward the evaluation of crew workload assessment techniques for aircraft certification. This study was conducted as a joint effort by the two major U.S. manufacturers of commercial transport airplanes: Douglas Aircraft Company and Boeing Commercial Airplanes. The primary purpose of this volume is to report the results of the contract effort. The objective of this contract was to provide assessment criteria to enable the FAA to evaluate workload measurement plans for crew size substantiation and workload acceptability during aircraft certification efforts, not to define a single measure or battery of measures that each manufacturer must use.

The authors wish to express appreciation to the many pilots from American, Delta, Eastern, TWA, and United Airlines who participated in this project. Extreme gratitude is also expressed to Preston Sult (Boeing Commercial Airplanes) for his participation as First Officer and differences trainer for all testing conducted. Thanks go out to Pat Kullenberg (Douglas Aircraft Company) for his help in debugging the flight scenarios.

The participation of flight engineers, Mike Bortolussi, Hugh Campion, and Doranne VonEnde is also appreciated. The efforts of the Manned Vehicle Simulation Research Facility (MVSRF) of NASA-Ames Research Center, and their simulation sub-contractors Singer-Link and Northrup in the preparation of the simulator and their help in conducting the tests are also appreciated. Thanks to Todd Williams (DAC) for his development of the Fortran programs used to reduce the simulator data. Gratitude is expressed to Vern Battiste and Sandra Hart of NASA-Ames Research Center for their participation in the part-task simulation. The efforts of Janet Camarata (BCA) in the preparation of the final manuscripts is greatly appreciated. Thanks also go to our contract monitors Harry Britten-Austin and Mike Pharaoh (RAF/USAF), and Peter Hwoschinsky (FAA) for their help and guidance throughout this program.

Thanks go to the personnel at the Man-Vehicle Systems Research Facility at NASA-Ames including: Elliott Smith, Rod Ketchum, David Gates, Barry Sullivan, Bob Shipley, Bob Shiner, and the cast and crew who feed the VAX all the 1's and 0's that make the simulation work.

Finally, the authors wish to thank the following student interns from California State University of Long Beach for their help in the data reduction and analyses: David Nixon, Mariel Sipman, Diana Kargoo, Toni LaFranchi, Terry Knight, Judy Wong, and Peggy Dolan.

SUMMARY

This report summarizes the work conducted as part of an FAA/U.S. Air Force sponsored contract (F33615-C-86-3600) "The Assessment of Crew Workload Measurement Methods, Techniques, and Procedures." The primary goal of the contract was to identify assessment techniques which demonstrate evidence of validity and reliability and are suitable as measures of flightcrew workload for aircraft certification.

To use a workload assessment technique with confidence for the certification of an aircraft flightdeck, the validity and reliability of the technique must be well established. Validity is the capability of the assessment technique to measure the abstract construct it is proposed to measure. Reliability is the capability of the measure to produce the same results with repeated testing.

A comprehensive literature review was conducted to identify workload measures which have an empirical record of validity and reliability. All candidate workload assessment techniques had to be applicable for evaluating workload in an aircraft environment. Two workshops were conducted to bring together experts in the workload assessment field to determine candidate measures for simulation testing (aided by the literature search), and make recommendations for testing in a high fidelity simulation. Two separate simulation tests were conducted at the Man Vehicle System Research Facility at NASA-Ames Research Center using a Phase II B-727 motion-base simulator.

The process by which this contract was conducted allows us to make factual statements regarding the validity and reliability of workload measures. The findings of validity and reliability for the workload measures tested are repeatable as demonstrated by the replication of results in the second simulation study. The method employed in this contract allows for an audit trail of the process by which an assessment technique is determined to be valid and reliable. A summary of the steps completed for this contract includes:

a. Literature review and Fact Matrices;
b. Workshop to gather expert agreement;
c. Simulation testing.

Workload measures which demonstrated evidence of validity and reliability in simulation testing includes:

a. In-flight and Post-flight subjective ratings (SWAT, NASA-TLX, and Bedford rating scales).
b. Heart rate, as measured by R-to-R wave Interbeat Interval.
c. Control Input Activity for the wheel (aileron) and column (elevator) during manual flight path control.

1.0 INTRODUCTION

Following the results of the 1981 Presidential Task Force on Crew Complement, much discussion has ensued regarding the process of certificating crew workload in commercial transport aircraft. The 1981 Task Force recommended improved workload assessment techniques be brought to bear on the certification process (McLucas, Drinkwater, and Leaf 1981). The contract (F33615-86-C-3600), reported in Volumes One and Two of this Final Report, was awarded to determine a list of acceptable workload measures for aircraft certification. The criteria which we propose should be employed for determining the representative workload measures to be included on the list were:

(a) Validity (The measure could quantitatively assess variations in workload.)

(b) Reliability (The measure would yield the same results with repeated application, over various crews.) This use of the word reliability is in reference to the statistical sense, and should not be confused with the usage as in reliability and maintainability.

(c) Applicability (The measure could be implemented in the transport flightdeck.)

No new workload measures were developed during the course of this contract. Existing workload measures were identified (literature search); selected based on demonstrated validity and reliability in empirical settings (literature search and first workshop); and evaluated empirically in the commercial transport environment (two simulation studies in a Phase II simulator at NASA-Ames). Volume One contains summaries of the two user-community workshops conducted and the empirical results from the simulation studies conducted at the Ames Research Center. Volume Two contains guidelines to be used by personnel evaluating a workload certification test plan.

The remaining sections of Volume Two contain:

(a) Guidelines which present preferred empirical practices for workload evaluation for aircraft certification,

(b) A subset of candidate workload measures for aircraft certification which demonstrated evidence of validity and reliability,

(c) Lessons learned regarding the utilization of the various workload measures listed in the implementation sections.

1.1 OBJECTIVES

The primary purpose of these guidelines is to enable the FAA and USAF to evaluate workload assessment plans. This is to be accomplished by providing:

(a) A list of measures which have exhibited evidence of validity and reliability in the assessment of civil transport workload,

(a) A list of measures which have exhibited evidence of validity and reliability in the assessment of civil transport workload,

(b) Methodology recommendations for implementation of the workload measures,

(c) A process for evaluating candidate workload measurement techniques and task scenarios.

Several specific objectives are identified to facilitate the evaluation of a workload certification plan and will provide:

(a) Guidelines for evaluating a proposed aircraft workload certification plan that will enable the FAA to insure that the workload criteria specified in FAR 15.2523, Appendix D are adequately considered,

(b) Guidelines for evaluating the adequacy of the proposed workload measures and scenarios included in a workload certification plan,

(c) Examples of evaluation criteria for the determination of acceptable workload measures and scenarios,

(d) Guidelines on how to evaluate the application of workload assessment techniques in aircraft certification,

(e) A data base to aid the FAA and USAF in location of factual information about workload measures suitable for aircraft certification.

1.2 BACKGROUND

Requirements to assess transport aircraft crew workload have developed as a means of assuring that the task-demands imposed on the aircrew will not exceed the crew's ability to respond to them in a safe and timely fashion. The criteria for assessment of crew workload are specified in FAR 25.1523, Appendix D of FAR 25.1523, and in FAR 25.771 (see Table 1.2-1).

TABLE 1.2-1. FEDERAL AVIATION REGULATION REQUIREMENTS

FAR 25.771 PILOT COMPARTMENT

(a) Each pilot compartment and its equipment must allow the minimum flight crew (established under 25.1523) to perform their duties without unreasonable concentration or fatigue.

FAR 25.1523 MINIMUM FLIGHTCREW

The minimum flight crew must be established so that it is sufficient for safe operation, considering - -

(a) The workload on individual crewmembers;

(b) The accessibility and ease of operation of necessary controls by the appropriate crewmember; and

(c) The kind of operation authorized under 25.1525.

The criteria used in making the determinations required by this section are set forth in Appendix D.

FAR 25 APPENDIX D

Criteria for determining minimum flight crew. The following are considered by the Agency in determining the minimum flight crew under 25.1523:

a. Basic workload function. The following basic workload functions are considered:

(1) Flight path control
(2) Collision avoidance
(3) Navigation
(4) Communications
(5) Operation and monitoring of aircraft engines and systems
(6) Command decisions

b. Workload factors. The following workload factors are considered significant when analyzing and demonstrating workload for minimum flight crew determination:

(1) The accessibility, ease, and simplicity of operation of all necessary flight, power and equipment controls, including emergency fuel shutoff valves, electrical controls, electronic controls, pressurization system controls, and engine controls.

TABLE 1.2-1. FEDERAL AVIATION REGULATION REQUIREMENTS
(Continued)

(2) The accessibility and conspicuity of all necessary instruments and failure warning devices such as fire warning, electrical system malfunction, and other failure or caution indicators. The extent to which such instruments or devices direct the proper corrective action is also considered.

(3) The number, urgency, and complexity of operating procedures with particular consideration given to the specific fuel management schedule imposed by center of gravity, structural and other considerations of an airworthiness nature, and to the ability of each engine to operate at all time from single tank or source which is automatically replenished if fuel is also stored in other tanks.

(4) The degree and duration of concentrated mental and physical effort involved in normal operation and in diagnosing and coping with malfunctions and emergencies.

(5) The extent of required monitoring of the fuel, hydraulic, pressurization, electrical, electronic, deicing, and other systems while enroute.

(6) The actions requiring a crewmember to be unavailable at his assigned duty station, including: observation of systems, emergency operation of any control, and emergencies in any compartment.

(7) The degree of automation provided in the aircraft systems to afford (after failure or malfunctions) automatic crossover or isolation of difficulties to minimize the need for flight crew action to guard against loss of hydraulic or electric power to flight controls or to other essential systems.

(8) The communications and navigation workload.

(9) The possibility of increased workload associated with any emergency that may lead to other emergencies.

(10) Incapacitation of a flight crewmember whenever the applicable operating rule requires a minimum flight crew of at least two pilots.

c. Kind of operation authorized. The determination of the kind of operation authorized requires consideration of the operating rules under which the airplane will be operated. Unless an applicant desires approval for a more limited kind of operation, it is assumed that each airplane certificated under this Part will operate under IFR conditions.

1.3 SCOPE

The purpose of the guidelines contained in this report is to aid the FAA in evaluating a proposed workload certification plan to determine if the applicant proposes methods which are valid, reliable, and applicable. If the proposed method is not on the list then it is the applicant's responsibility to provide data demonstrating validity, reliability, and applicability.

Candidate workload measures for aircraft certification are provided as examples of valid and reliable assessment techniques. The measures listed in this volume are not intended to be an exclusive list of assessment techniques available, rather the candidate measures should serve as references for the reader.

There are several reasons that a list of acceptable measures could give misleading guidance. A partial listing would be:

(a) A measure was erroneously included on the list.

(b) The interaction of the measures on the list might evaluate the test conditions for one set of task demands adequately, but would not evaluate another set of task demands very well.

(c) The measure is not valid or reliable with the type of task demands it has been selected to index.

Other concerns to keep in mind are that:

(a) One measure cannot be employed to assess every type of workload.

(b) Only measures that assess workload in a relative fashion are currently available.

(c) The list of acceptable measures should not be limited to only those assessment techniques which are currently identified. New measures continue to be developed and research is continuously being conducted on existing techniques.

Analytic assessment techniques, along with subjective judgements, are the de facto standard for assessing workload. There are times when additional information, such as objective data, can be of benefit understanding workload based on complex task demands in the aircraft environment. The best role for objective data may be to supplement analytic and subjective measures. Certainly, the existing process of using timeline analysis (TLA) to anticipate and support subjective assessments has been useful and is associated with aircraft which continue to have excellent safety records.

1.3.1 PROGRAM CONSTRAINTS

Owing to the large number of possible workload measures which could be reviewed and evaluated, certain self-imposed limitations were outlined to insure adequate resources would be available for a reasonable quality simulation

evaluation of the candidate measures. The following limitations were outlined:

(a) Candidate workload measures were selected from those which already existed. A candidate measure had to have published evidence of validity and reliability.

(b) The representative measures were evaluated from each of three categories: Subjective, Physiological, and Performance workload assessment techniques. Measures were chosen based on exhibiting the most evidence that they were valid, reliable, and applicable to aircraft certification. The number of measures to be evaluated was limited by the available resources.

(c) Only measures which were suitable for full fidelity simulation or flight test were evaluated.

(d) Measures were evaluated in a civil transport environment (as opposed to military environments). The types of task demands addressed in scenario selection were identified by, but not limited to, the functions and factors in FAR 25.1523, Appendix D. (The results obtained from scenarios which are common with military task demands will be valid for military applications.)

(e) The issues of underload and fatigue were not examined.

(f) Military standards were incorporated in these guidelines where possible.

(g) A subset of acceptable workload measures is included.

The work reported in this Final Report was never intended to develop a battery of measures for flightdeck workload certification. That approach would presuppose that the task demands for all transport aircraft are the same, and could be addressed by a standardized set of assessment techniques. In fact workload certification for one aircraft may focus on handling qualities where as another may focus on visual information processing with electronic displays. In addition, the battery would need to be tested for validity and reliability, as opposed to individual workload measures as was done in simulation studies.

If measures are developed and validated as a test battery, then it cannot be expected that they reflect independent (unique) measures of workload.

1.3.2 CERTIFICATION APPLICATIONS

The most relevant consideration when evaluating a workload assessment plan is that the proposed workload techniques are adequate for evaluating the anticipated workload, or workload changes, for the new flight deck.

In the past, commercial aircraft manufacturers have used analytic techniques and non-structured pilot opinion for workload assessment. Analytic techniques are of particular value to the aircraft manufacturer since they offer both the potential for identifying and correcting workload problems early in the design phase when the cost of change is relatively low, and a tool which can provide data for certification.

One disadvantage to the available analytic techniques is their lack of fidelity in assessing mental effort. With the current shift of flight deck design placing more mental demands on the flight crew, workload assessment has taken on a new challenge. The addition of structured subjective measures to traditional objective analyses can provide information which validates the analytic and simulation based estimates of physical workload and enhances estimates of mental workload.

An evaluation plan should not only consider whether the appropriate assessment techniques are used for answering questions of workload, but also whether the techniques are used correctly. A partial listing of common errors is as follows:

(a) Confounding differences between aircraft and evaluation conditions - Workload evaluation conditions should be identical between the baseline and to-be-certificated aircraft.

(b) Using the same order of evaluation conditions introducing systematic biases into the workload assessment process - An example is placing all the baseline aircraft simulation runs ahead of the to-be-certificated aircraft on a given day. This could introduce fatigue as an uncontrolled variable in the evaluation process.

(c) Introducing bias into the subjective rating process - For example, hints are inadvertently given to the subject on how to rate the workload, either high or low.

1.3.3 CONSIDERATION OF MILITARY STANDARDS

A review was made of the military standard entitled: Human Engineering Requirements for Military Systems, Equipment, and Facilities (MIL-H-46855A) and the draft military standard entitled: Human Engineering Requirements for Measurement of Operator Workload. Methodology applicable to commercial transport aircraft certification was incorporated into the study.

1.3.4 RELATIVE VERSUS ABSOLUTE MEASUREMENT

Workload assessment for certification relies on a relative comparison of workload levels. Typically, workload is compared between the to-be-certificated aircraft and a baseline aircraft, which has an established record of safe performance and acceptable workload. We assumed that the two aircraft are being compared under conditions which are as similar as possible to insure that any workload differences which occur are due to differences in the aircraft design and not to other factors (requirement for valid experimental design). If the new model aircraft has the same or lower workload, then it is concluded that the workload is acceptable in the new model. When performing a relative comparison with a new aircraft design, however, there may be instances when workload levels exceed the old design. In cases such as this, the increased workload is not necessarily unacceptable, but it may become the subject of a more in-depth workload analysis. These cases need to be considered on a case-by-case basis with all of the operational factors taken into consideration when evaluating the impact of small workload increases.

When measuring pilot workload, or any other behavioral measure, it is essential to consider the variable nature of the data. Behavioral data is best described in terms of distributions, since individuals bring different skills to the task of flying it is possible to get a distribution of workload scores from a group of pilots. The state of the art of workload science does not allow for determination of a single score for the purpose of workload assessment. Pilot to pilot variability in assessing workload is a consideration which must be kept in mind throughout an aircraft certification effort. No absolute measure of workload is currently possible for aircraft certification.

A number of factors influence the ability to generalize or draw conclusions about workload levels made in a comparative evaluation. It would not be appropriate to include a detailed discussion of these factors here, but a partial listing of relevant factors includes:

(a) Representativeness of subject selection

(b) Number of subjects tested

(c) Fidelity of task demands or scenarios

(d) Adequate content validity (e.g., inclusion of relevant workload types and critical mission segments).

1.4 CURRENTLY USED TECHNIQUES

Today's list of acceptable workload measures is not likely to be tomorrow's list. Any list which is fixed and cannot be modified to accommodate the improvements developed within the workload measurement science, could become more of an obstacle than an aid in certifying the design of a new aircraft. For this reason, emphasis should be placed on whether the most useful measure was selected for a particular application, rather than selection of a measure merely because it was familiar and associated with an established list.

2.0 CHECKLIST FOR WORKLOAD EVALUATION

The following section is an abbreviated version of sections 3, 4, and 5. The purpose of this section is to provide for the person evaluating a workload certification test plan to determine if the test plan complies with empirically sound practices and relevant lessons learned included in the implementation sections.

This section should not be construed as a binding requirement for a workload certification effort. If, in the course of evaluating a workload certification test plan, certain areas are deemed weak, the following guidelines can serve as a vehicle for a discussion of the test plan.

2.1 VALIDITY - RELIABILITY - APPLICABILITY GUIDELINES

2.1.1 VALIDITY GUIDELINES

CONSTRUCT VALIDITY

- o EVIDENCE SHOULD BE PROVIDED TO SHOW THAT THE WORKLOAD ASSESSMENT TECHNIQUE REFLECTS CHANGES IN OPERATIONALLY RELEVANT TYPES OF WORKLOAD.

FACE VALIDITY

- o WHENEVER POSSIBLE, THE RELEVANCE OF THE WORKLOAD ASSESSMENT TECHNIQUE SHOULD BE EXPLAINED AND ILLUSTRATED TO SHOW HOW IT IS A MEASURE OF WORKLOAD.

2.1.1.1 VALIDITY EVALUATION CRITERIA

- o ASSESSMENT TECHNIQUES USED FOR EVALUATING FLIGHT DECK WORKLOAD SHOULD HAVE DEMONSTRATED THE ABILITY TO DISCRIMINATE WORKLOAD FUNCTIONS AND FACTORS IDENTIFIED IN FAR 25.1523, APPENDIX D.

2.1.2 RELIABILITY GUIDELINES

TEST-RETEST RELIABILITY

- o THE WORKLOAD ASSESSMENT TECHNIQUE SHOULD PROVIDE SIMILAR RESULTS WITH REPEATED TESTING OF THE SAME PILOTS.

INTER-RATER RELIABILITY

- o THE WORKLOAD ASSESSMENT TECHNIQUE SHOULD BE RELATIVELY STABLE ACROSS DIFFERENT PEOPLE. FOR EXAMPLE, THOUGH THE ACTUAL SCORES MAY BE DIFFERENT THE RESULTS FROM VARIOUS PILOTS SHOULD HAVE A SIMILAR APPEARANCE (HIGH SCORES CLUSTER FOR THE HIGH WORKLOAD CONDITIONS AND LOW SCORES CLUSTER FOR THE LOW WORKLOAD CONDITIONS).

2.1.2.1 RELIABILITY EVALUATION CRITERIA

- o ASSESSMENT TECHNIQUES USED FOR EVALUATING FLIGHT DECK WORKLOAD SHOULD HAVE A DEMONSTRATED RECORD OF REPEATABILITY.

2.1.3 APPLICABILITY GUIDELINES

APPLICABILITY

- o THE ASSESSMENT METHOD SHOULD BE APPROPRIATE FOR THE SPECIFIC PHASE AND OBJECTIVES OF THE CERTIFICATION PROGRAM.

OBTRUSIVENESS

- o THE ASSESSMENT METHOD SHOULD CAUSE MINIMAL INTERFERENCE WITH OTHER CERTIFICATION FLIGHT TEST ACTIVITIES AND NOT SIGNIFICANTLY CHANGE THE WORKLOAD OF THE CREW.

CREW SAFETY

- o THE ASSESSMENT METHOD SHOULD BE NON-INTERFERING WITH REGARD TO FLYING ACTIVITIES.

CAREER THREAT

- o THE APPLICATION OF THE ASSESSMENT METHOD SHOULD BE NON-CAREER THREATENING TO THE CREW MEMBERS IT EVALUATES. FOR EXAMPLE, PHYSIOLOGICAL MEASURES SHOULD GIVE NO DIAGNOSTIC MEDICAL INFORMATION.

CERTIFICATION INTERFACE

- o WORKLOAD ASSESSMENT TECHNIQUES WHICH ARE TIME-INTENSIVE AND RISK DELAYS IN CERTIFICATION SCHEDULES SHOULD BE AVOIDED THAT .

FLIGHT ENVIRONMENT CONSTRAINTS

- o THE ASSESSMENT METHOD SHOULD BE CAPABLE OF GATHERING DATA UNDER THE CONSTRAINTS OF THE FLIGHT TEST OR HIGH FIDELITY PILOTED SIMULATION.

2.2 WORKLOAD ASSESSMENT GUIDELINES

- o A WORKLOAD ASSESSMENT TECHNIQUE USED FOR AIRCRAFT CERTIFICATION SHOULD EXHIBIT EVIDENCE THAT IT IS VALID, RELIABLE, AND APPLICABLE.

- o WORKLOAD ASSESSMENT TECHNIQUES USED FOR AIRCRAFT CERTIFICATION SHOULD ADDRESS SPECIFIC FUNCTIONS AND FACTORS OF FAR 25.1523, APPENDIX D.

- o DATA SHOULD BE COLLECTED IN MEASUREMENT WINDOWS THAT ARE COMPARABLE FOR ALL THE WORKLOAD MEASURES USED IN THE CERTIFICATION EFFORT.

2.3 CANDIDATE SUBJECTIVE WORKLOAD MEASURES FOR AIRCRAFT CERTIFICATION

- The Bedford (Modified Cooper-Harper type), the Subjective Workload Assessment Technique (SWAT), and the NASA Task Load Index (TLX) have demonstrated evidence of validity, reliability, and applicability as measures for evaluating pilot subjective workload post-flight.

- Although not evaluated in the simulation studies at NASA-Ames, comparative subjective evaluation techniques (e.g., Pilot Subjective Evaluation-PSE) have previously demonstrated applicability for directly comparing two different aircraft, such as a baseline and a new aircraft.

2.3.1 IN-FLIGHT SUBJECTIVE MEASURE IMPLEMENTATION GUIDELINES

- o PILOTS USED FOR THE WORKLOAD ASSESSMENT SHOULD BE UNAWARE OF THE MANIPULATION OF TASK DEMANDS (MALFUNCTIONS, DIVERSIONS, ETC.) DURING THE EVALUATION FLIGHTS.

- o WHEN USING A SUBJECTIVE MEASURE IN-FLIGHT, THE MEASURE SHOULD NOT BE INTRUSIVE TO THE FLIGHT RELATED TASKS THE CREW MEMBER IS TRYING TO ACCOMPLISH.

- o IF PAPER AND PENCIL RATING TECHNIQUES ARE TO BE USED IN FLIGHT, ONE CREW MEMBER AT A TIME SHOULD RECORD THEIR WORKLOAD RATINGS SO THAT THE OTHER CREW MEMBER MAY ATTEND TO FLIGHT DECK DUTIES.

- o THE TO-BE-RATED FLIGHT SEGMENT (BEGINNING AND END POINTS) SHOULD BE CLEARLY IDENTIFIED TO THE FLIGHT CREW FOR THE PURPOSE OF OBTAINING ACCURATE DATA FOR EVALUATION.

2.3.2 POST-FLIGHT SUBJECTIVE MEASURE IMPLEMENTATION GUIDELINES

- o TO ENHANCE POST-FLIGHT WORKLOAD EVALUATION, VIDEOTAPE SHOULD BE USED TO AID THE CREW IN

RECALLING THEIR SUBJECTIVE EVALUATIONS OF CREW WORKLOAD.

- o THE TO-BE-RATED FLIGHT SEGMENT (BEGINNING AND END POINTS) SHOULD BE CLEARLY IDENTIFIED TO THE FLIGHT CREW FOR THE PURPOSE OF OBTAINING ACCURATE DATA FOR EVALUATION.

- o WHEN USED, POST-FLIGHT SUBJECTIVE RATINGS SHOULD BE COLLECTED FROM THE PILOTS AS SOON AFTER THE TASK AS OPERATIONALLY FEASIBLE.

2.4 CANDIDATE PHYSIOLOGICAL WORKLOAD MEASURES FOR AIRCRAFT CERTIFICATION

- Heart rate over a period of time (in beats per minute or Inter-Beat Interval) has demonstrated evidence of reliability as a measure of pilot workload.

2.4.1 PHYSIOLOGICAL IMPLEMENTATION GUIDELINES

- o DATA COLLECTED WITH PHYSIOLOGICAL MEASURES CAN BE CONTAMINATED BY PHYSICAL MOVEMENT. SOURCES OF ARTIFACTS SHOULD BE CONTROLLED WHEN EVALUATING THE IMPLEMENTATION OF A WORKLOAD MEASURE.

- o THE DATA SHOULD BE REPRESENTATIVE OF THE ENTIRE FLIGHT SEGMENT BEING EVALUATED, SO SOME SORT OF AVERAGING SHOULD BE USED WITHIN THE FLIGHT SEGMENT.

- o CARE SHOULD BE TAKEN SO THAT THE FLIGHT CREW IS PROTECTED FROM HAZARDS, SUCH AS ELECTRICAL SHOCK.

- o CARE SHOULD BE TAKEN TO ASSURE THAT THE PHYSIOLOGICAL ASSESSMENT METHOD APPEARS NON-CAREER THREATENING TO THE CREW MEMBERS IT EVALUATES (E.G., DATA COLLECTED USING PHYSIOLOGICAL MEASURES SHOULD CONTAIN NO DIAGNOSTIC MEDICAL INFORMATION).

2.5 CANDIDATE PERFORMANCE (PRIMARY TASK) WORKLOAD MEASURES FOR AIRCRAFT CERTIFICATION

- Control input activity (i.e., wheel, column, and pedal) has demonstrated evidence of validity, reliability, and applicability as performance measures for evaluating pilot workload.

2.5.1 PRIMARY TASK IMPLEMENTATION GUIDELINES

- o CONTROL INPUT ACTIVITY SHOULD BE EVALUATED ONLY DURING MANUAL FLIGHT PATH CONTROL.

- o WHEN POSSIBLE, STATE VARIABLES (E.G., PITCH ANGLE, ROLL ANGLE, ALTITUDE) SHOULD BE RECORDED CONTINUOUSLY IN SIMULATION TESTS.

- o WHEN POSSIBLE, WHEEL (AILERON) AND STICK (ELEVATOR) INPUTS SHOULD BE EMPLOYED TO REPRESENT AIRCRAFT CONTROL WORKLOAD THROUGHOUT THE ENTIRE FLIGHT OF AN AIRCRAFT UNDER MANUAL FLIGHTPATH CONTROL.

- o PEDAL (RUDDER) ACTIVITY IS NORMALLY ONLY REPRESENTATIVE OF AIRCRAFT CONTROL IN THE TAKEOFF AND APPROACH/LANDING PHASE OF THE FLIGHT AND SHOULD BE COLLECTED DURING THESE FLIGHT PHASES.

- o THE SAME FLIGHT SCENARIO SHOULD BE USED WHEN COMPARING NEW AND BASELINE AIRCRAFT.

- o A FLIGHT SHOULD BE DIVIDED INTO SEGMENTS FOR DATA COLLECTION SO DESCRIPTIVE STATISTICS (E.G., WHEEL POSITION, COLUMN POSITION) CAN BE COMPUTED ON THE CONTINUOUS MEASURES WITHIN EACH SEGMENT.

2.5.2 SECONDARY TASK IMPLEMENTATION GUIDELINES

- o WHEN USED, SECONDARY TASKS SHOULD BE EMBEDDED IN THE FLIGHT TASK SO AS TO BE AS NON-INTRUSIVE AS POSSIBLE.

- o EMBEDDED SECONDARY TASKS SHOULD NOT APPEAR ARTIFICIAL TO THE OPERATOR SO AS TO MAINTAIN OPERATOR ACCEPTANCE AND FACE VALIDITY.

- o SECONDARY TASKS ARE MOST EFFECTIVELY IMPLEMENTED IN A SIMULATION ENVIRONMENT, WHERE AIR SAFETY IS NOT A CONCERN AND CONTROL OF THE ENVIRONMENT IS POSSIBLE.

- o SECONDARY TASK TECHNIQUES SHOULD BE AVOIDED WHEN INTRUSION WILL SERVE AS A SOURCE OF INTERFERENCE FOR THE PRIMARY WORKLOAD MEASURES.

2.6 CANDIDATE ANALYTIC ASSESSMENT TECHNIQUES FOR AIRCRAFT CERTIFICATION

- The analytic assessment technique of Timeline Analysis has demonstrated evidence of validity and applicability for assessing crew task demands.

2.6.1 ANALYTIC TECHNIQUES IMPLEMENTATION GUIDELINES

- o WHEN USING ANALYTIC TECHNIQUES, CONCLUSIONS REGARDING WORKLOAD ACCEPTABILITY SHOULD BE BASED ON RELATIVE WORKLOAD COMPARISONS BETWEEN THE TO-BE-CERTIFICATED AIRCRAFT AND BASELINE.

- o WHEN PERFORMING A RELATIVE COMPARISON WITH A NEW AIRCRAFT DESIGN, AND WORKLOAD LEVELS EXCEED THE OLD DESIGN FOR A FLIGHT SEGMENT, A DECISION SHOULD BE MADE REGARDING THE NEED FOR A MORE IN-DEPTH WORKLOAD ASSESSMENT.

- o DETAILED PROCEDURES SHOULD BE DEVELOPED THAT DEFINE ALL ACTIONS EACH CREW MEMBER MUST ACCOMPLISH TO COMPLETE A FLIGHT SUCCESSFULLY.

- o CREATION OF THE SCENARIO SHOULD BE BASED UPON DATA DERIVED FROM FLIGHT PLANS, NAVIGATION CHARTS (SID, STAR, ENROUTE AREA, APPROACH, AND AIRPORT MAPS), ATC OPERATIONAL DATA, AIRCRAFT PERFORMANCE DATA, AND AIRCRAFT OPERATIONS MANUALS.

2.7 SCENARIO DESCRIPTION GUIDELINES

- o DISCRETE MEASUREMENT PERIODS SHOULD BE USED FOR EVALUATING WORKLOAD, OTHERWISE SPECIFIC EVENTS OR ACTIONS ARE MIXED WITH UNWANTED TYPES OF WORKLOAD IRRELEVANT TO THE CONCERNS FOR CERTIFICATION. THIS LEADS TO THE INABILITY TO EVALUATE DISCRETE VARIATIONS IN CREW WORKLOAD.

2.7.1 SCENARIO EVALUATION CRITERIA

ROUTE CHOICE

- o THE ROUTES SHOULD PROVIDE A REPRESENTATIVE MIX OF NAVIGATION AIDS, AIRPORTS, APPROACHES, AND AIR TRAFFIC CONTROL (ATC) SERVICES. IN ADDITION, ROUTES THAT ADEQUATELY SAMPLE HIGH DENSITY TRAFFIC AREAS.

NEW TECHNOLOGY AND EQUIPMENT

- o THE SCENARIO FLOWN SHOULD TAKE INTO ACCOUNT THE NEW EQUIPMENT INCORPORATED INTO THE TO-BE-CERTIFICATED AIRCRAFT. PROCEDURES ASSOCIATED WITH THE NEW EQUIPMENT SHOULD BE ADDRESSED, AS WELL AS OPERATIONAL AND MINIMUM EQUIPMENT LISTS.

CRITICAL EVENT IDENTIFICATION

- o THE SCENARIO FLOWN SHOULD REPRESENT THE RANGE OF OPERATIONAL REQUIREMENTS OF THE TO-BE-CERTIFICATED AIRCRAFT INCLUDING REPRESENTATIVE NORMAL AND NON-

NORMAL PROCEDURES LIKELY TO BE ENCOUNTERED DURING ITS OPERATION IN SERVICE.

PROCEDURES USED TO RELATE WORKLOAD TO FAR REQUIREMENTS

- WORKLOAD FUNCTIONS AND WORKLOAD FACTORS DESCRIBED IN FAR 25 APPENDIX D SHOULD BE REPRESENTED IN THE SCENARIOS FLOWN IN THE CERTIFICATION EFFORT.

3.0 EVALUATION CRITERIA FOR ASSESSMENT TECHNIQUES

The guidelines in this section are general, and apply to the evaluation of workload assessment techniques proposed for aircraft certification. Airframe manufacturers should use assessment techniques that are determined to be valid (the measure can assess workload in a quantitative fashion), reliable (the measure produces the same results with repeated application) and applicable (the measure can be used in a full fidelity flightdeck environment).

3.1 VALIDITY GUIDELINES

CONSTRUCT VALIDITY

- EVIDENCE SHOULD BE PROVIDED TO SHOW THAT THE WORKLOAD ASSESSMENT TECHNIQUE REFLECTS CHANGES IN OPERATIONALLY RELEVANT TYPES OF WORKLOAD.

FACE VALIDITY

- WHENEVER POSSIBLE, THE RELEVANCE OF THE WORKLOAD ASSESSMENT TECHNIQUE SHOULD BE EXPLAINED AND ILLUSTRATED TO SHOW HOW IT IS A MEASURE OF WORKLOAD.

The question should be asked; Does the workload assessment technique really measure what it is intended to measure? There are many types of validity, each affecting the ultimate usefulness and acceptability of a workload measure. At a minimum it is proposed that construct and face validity be addressed in every stage of measure selection, testing, and evaluation (Anastasi, 1968).

All operationally relevant types of workload should be considered when considering a given workload measure. By focusing on the significant types of workload found in cockpit operations, more confidence can be obtained that the correct workload assessment techniques will be selected and employed.

The construct validity of a workload assessment technique is the extent to which the technique may be said to measure the theoretical construct of workload. Since workload cannot be directly observed, it exists only as an abstract concept, it must be demonstrated that the measure in fact reflects changes in workload. To have confidence in a workload measure, this connection must be demonstrated whether workload is defined in terms of task demands or operator variables. Construct validity is not accomplished in a single experiment or settled once and for all, it requires the gradual accumulation of information from a variety of sources.

Face validity refers to what the assessment technique appears superficially to measure and not to what it actually measures. Face validity can become important in how well people use an assessment technique. If pilots or engineers are asked to administer a workload measurement system which makes little sense to them, their motivation to follow all the rules is likely to suffer. The results of an invalid application of a good measure can be worse than no measure at all.

3.1.1 VALIDITY EVALUATION CRITERIA

- o ASSESSMENT TECHNIQUES USED FOR EVALUATING FLIGHT DECK WORKLOAD SHOULD HAVE DEMONSTRATED THE ABILITY TO DISCRIMINATE WORKLOAD FUNCTIONS AND FACTORS IDENTIFIED IN FAR 25.1523, APPENDIX D.

Workload assessment techniques employed in new aircraft certification should have a demonstrated ability to discriminate levels of workload, as defined operationally bY FAR 25.1523, Appendix D functions and factors. Content, construct, and face validity are all concepts which should be considered when selecting assessment techniques for evaluating workload. Workload measures should have a demonstrated ability to address these issues before being used for aircraft certification.

Various methods demonstrate discriminability among levels of workload. Previous methods used in aircraft certification (i.e., Timeline Analyses) have a proven record for the assessment of physical workload. The ability to discriminate between flights where a single workload function or factor has been varied would provide evidence of validity. Different phases of flight, within the same flight, require different Appendix D functions and factors for flying the aircraft. A new workload technique should have an empirical record of discriminability for flights exercising different FAR 25.1523, Appendix D, functions and factors, as well as discriminating different phases of flight from one another.

Evidence of validity can be found in many forms, the most notable is literature published which demonstrates validity from an empirical investigation. Airframe manufacturers do not have the resources necessary to investigate each new measure of workload. The time tested tradition of citing published results from scientific journals is an acceptable method for determining the validity of workload measures. It should be expected that the results published are in some fashion connected with the application of the workload measure in an aviation environment. Tests over varying conditions which replicate the finding of discriminability of different workload functions and factors by an assessment technique have the strongest evidence for validity.

3.2 RELIABILITY GUIDELINES

TEST-RETEST RELIABILITY

- o THE WORKLOAD ASSESSMENT TECHNIQUE SHOULD PROVIDE SIMILAR RESULTS WITH REPEATED TESTING OF THE SAME PILOTS.

INTER-RATER RELIABILITY

- o THE WORKLOAD ASSESSMENT TECHNIQUE SHOULD BE RELATIVELY STABLE ACROSS DIFFERENT PEOPLE. FOR EXAMPLE, THOUGH THE ACTUAL SCORES MAY BE DIFFERENT THE RESULTS FROM VARIOUS PILOTS SHOULD HAVE A SIMILAR APPEARANCE (HIGH SCORES CLUSTER FOR THE HIGH WORKLOAD CONDITIONS AND LOW SCORES CLUSTER FOR THE LOW WORKLOAD CONDITIONS).

An important concern is the reliability of the workload assessment technique. How consistently does the measure yield the same answer given the same measurement conditions? Certainly a workload assessment technique must be highly reliable (repeatable) before its results become the basis for design decisions. It is proposed that the following types of reliability be addressed (Anastasi, 1968): (a) test-retest reliability, and (b) inter-rater reliability.

An unreliable workload measure can create problems for either the flight crew or the manufacturer by yielding one of two possible types of errors. The first error (type I) occurs when the measure indicates that the workload is excessive, when in fact it is not. This type of error during certification could impose significant costs on manufacturers and purchasers of aircraft. The results of a second type of error are also unacceptable. This error (type II) occurs when the measure indicates that the workload is acceptable, when in fact it is not. This type of error could lead to the sale and operation of less acceptable aircraft. Since there is the risk of making either type of error, only the most reliable measures should be employed.

Although it is convenient to think of people as experiencing a similar level of workload in response to a set of fixed task-demands (e.g., an average workload level), constant workload levels cannot be assigned because of the individual nature of each person's actions (Hart and Bortolussi, 1983). Not only are people widely different in factors which determine the difficulty of a task (e.g., training, skill), but the workload level experienced by an individual can vary widely over time due to fatigue or health factors, even in response to constant task-demands (Damos, 1984). The same individual can be expected to give different ratings of workload, in response to identical task demands, on different days. The full fidelity simulation study reported in Volume One of this Final Report provides test-retest reliability results for workload measures in a commercial aircraft transport environment.

Test-retest methods for establishing the reliability of an assessment method is to repeat the identical assessment on a second occasion. The reliability can be quantified by computing the correlation between the two sets of scores obtained by the same pilots on the two administrations of the workload measure. The resulting reliability coefficient can then be compared one to another for any test, and thereby be viewed with some objectivity. If an assessment technique is used to discriminate between high and low levels of workload, it should discriminate between high and low levels the same way on a second occasion.

Inter-rater reliability is a method employed to determine the consistency of an assessment technique across different people. One technique for assessing inter-rater reliability is to correlate each pilot's workload scores, for a variety of conditions, with the group mean of the conditions in the same test.

3.2.1 RELIABILITY EVALUATION CRITERIA

- o ASSESSMENT TECHNIQUES USED FOR EVALUATING FLIGHT DECK WORKLOAD SHOULD HAVE A DEMONSTRATED RECORD OF REPEATABILITY.

Workload assessment techniques employed in new aircraft certification should have demonstrated the ability to produce similar results with repeated application to the same, and various flight crews.

A workload measure which yields one set of results on initial application, but a contradictory set of results with repeated application is considered unreliable. Proven test-retest reliability of a workload assessment technique is a necessary condition for the consideration of the technique for aircraft certification. With greater experience with an aircraft's operations, perceived workload often decreases. A decrease in overall workload with experience does not effect the evaluation of test-retest reliability. If the same pattern of low and high workload is consistent, but generally lower for subsequent flights, then test-retest reliability will be high.

Inter-rater reliability is a difficult concept for workload assessment techniques to adequately address. It is important for a workload assessment technique to be generalizable to the entire population of pilots who will later fly the aircraft.

Sophisticated workload researchers have sought stability of workload measures, as opposed to reliability. If a workload measure yields the same results for an entire group, whose training and experience differs, then this is thought to be a liability by workload assessment experts, because of the measure's insensitivity to individual differences. A workload measure should reflect the difference among individuals based on training and experience. Yet, stability and inter-rater reliability are not mutually exclusive concepts.

For inter-rater reliability coefficients to be high it is required that the same low and high task demands yield low and high workload assessments across pilots be independent of the absolute value. For example, a less experienced pilot may render workload values of "50," "25," and "75" for different flight phases, while the experienced test pilot renders values of "15," "7," and "25" for the same flight phases, inter-rater reliability, as well as stability, are found to be high. The reliability is high because the order of difficulty is the same although the absolute values of the ratings are different.

A completely satisfactory method for computing stability of a workload measure across subjects has not been defined. It is recognized that a need exists for improving methods for computing the stability of a measure. It is beyond the scope of this work to go into an in depth discussion how new methods could be developed to address stability. The method for addressing stability in this contract, namely test-retest and inter-rater reliability, is consistent with recommended statistical methods (Anastasi, 1968).

New methods for workload assessment are continually being developed. As was the case for validity, evidence of the reliability of a workload measure can be found in many forms, the most notable is literature published which demonstrates reliability from an empirical investigation. It should be expected that the results published are in some fashion connected with the application of the workload measure in an aviation environment. Studies employing a test-retest methodology, or comparing subject's scores to the group average, are one line of evidence for the reliability of candidate workload measures for aircraft certification. It is difficult to establish concrete criteria for whether or not an assessment technique is reliable. One straightforward criteria is whether the correlation (test-retest or inter-rater reliability) coefficients are significant, in other words the

relationship is not due to chance. At a minimum the correlations should be significant to provide evidence of reliability.

3.3 APPLICABILITY GUIDELINES

APPLICABILITY

- o THE ASSESSMENT METHOD SHOULD BE APPROPRIATE FOR THE SPECIFIC PHASE AND OBJECTIVES OF THE CERTIFICATION PROGRAM.

OBTRUSIVENESS

- o THE ASSESSMENT METHOD SHOULD CAUSE MINIMAL INTERFERENCE WITH OTHER CERTIFICATION FLIGHT TEST ACTIVITIES AND NOT SIGNIFICANTLY CHANGE THE WORKLOAD OF THE CREW.

CREW SAFETY

- o THE ASSESSMENT METHOD SHOULD BE NON-INTERFERING WITH REGARD TO FLYING ACTIVITIES.

CAREER THREAT

- o THE APPLICATION OF THE ASSESSMENT METHOD SHOULD BE NON-CAREER THREATENING TO THE CREW MEMBERS IT EVALUATES. FOR EXAMPLE, PHYSIOLOGICAL MEASURES SHOULD GIVE NO DIAGNOSTIC MEDICAL INFORMATION.

CERTIFICATION INTERFACE

- o WORKLOAD ASSESSMENT TECHNIQUES WHICH ARE TIME-INTENSIVE AND RISK DELAYS IN CERTIFICATION SCHEDULES SHOULD BE AVOIDED.

FLIGHT ENVIRONMENT CONSTRAINTS

- o THE ASSESSMENT METHOD SHOULD BE CAPABLE OF GATHERING DATA UNDER THE CONSTRAINTS OF THE FLIGHT TEST OR HIGH FIDELITY PILOTED SIMULATION.

The applicability of a workload measure to the flight deck environment is central to the entire workload assessment process. The workload assessment technique must apply to the types of workload which occur in aircraft systems relevant. In order for a workload measure to be applicable in a flight operational environment, it should satisfy a number of requirements such as those listed below:

(a) The assessment method chosen should not endanger the crew's safety nor interfere with their normal duties.

(b) To promote pilot acceptance, the assessment method should not impose an additional workload on the crew and thereby disturb the very process that it is trying to measure.

(c) In order to promote pilot acceptance, it is important that the assessment technique does not seem career threatening to the crew. For instance, data collected using physiological measures should contain no diagnostic medical information. Care should be taken to explain the purposes, procedures, and limits of all assessment techniques so as to limit operator apprehension and suspicion of their use.

Crew workload assessment is only a small part of the whole certification process. Because the certification process is time consuming and costly, many certification flight test activities are performed simultaneously. To be practical addition to any overall certification activity, it is very important that the workload assessment technique chosen should cause minimal interference with the many other activities that will be performed concurrently. To determine practicality of a given technique in aircraft systems the following considerations should be made:

(a) The less complicated the workload assessment process is, the better it will be understood by those individuals involved. This means it will have greater acceptance by the crew members it evaluates and less likely to be misinterpreted.

(b) Because the assessment technique must work in the real world, it must be able to handle several existing problems in the operational environment for the use of additional equipment. Evaluation of equipment constraints in the flight deck should give consideration to:

(1) Limited hardware and panel space,
(2) Impact on crew behavior due to changes in hardware,
(3) A large distance between pilot and data collection hardware,
(4) Potential signal interference.

(c) There are time constraints of the certification program schedule, production schedules, and delivery schedules. An unnecessary delay in any of these schedules is costly and unacceptable to both the manufacturers and their customers. Significant financial risks can destroy the very industry we are trying to help. The workload assessment technique, therefore, should be able to be completed within these time constraints.

(d) The costs incurred by the workload assessment techniques vary depending on their complexity and equipment requirements. Evaluation of the costs incurred should include:

(1) Equipment costs,
(2) Installation and preparation costs,
(3) Time and schedule impact,
(4) Flight and simulation costs, and
(5) Documentation costs.

It is important to consider the operational environment of the airplane when choosing a workload assessment technique. Confidence should be provided that the assessment method works in the real world and that it is a cost-effective procedure.

4.0 WORKLOAD ASSESSMENT TECHNIQUES

Workload assessment techniques are currently incapable of measuring overload across a population of pilots. Workload, as a science, has not perfected measuring pilot workload in an absolute sense, largely because of the profound effect of individual differences. Pilots with more experience must be loaded with many tasks before there is a breakdown in performance. Inexperienced pilots, with fewer total hours or fewer hours in aircraft type, become overloaded with much fewer tasks. The problem of workload assessment is therefore a problem of relative comparison. Workload assessment is a relative effort, in which an attempt is made to evaluate workload in direct comparison to a baseline.

Engineering concepts are traditionally expressed in terms of exact quantities and their associated tolerances. Workload assessment is better represented by a distribution of scores around an exact quantity such as an average score.

4.0.1 WORKLOAD ASSESSMENT GUIDELINES

- o WORKLOAD ASSESSMENT TECHNIQUES USED FOR AIRCRAFT CERTIFICATION SHOULD EXHIBIT EVIDENCE THAT IT IS VALID, RELIABLE, AND APPLICABLE.

- o WORKLOAD ASSESSMENT TECHNIQUES USED FOR AIRCRAFT CERTIFICATION SHOULD ADDRESS SPECIFIC FUNCTIONS AND FACTORS OF FAR 25.1523, APPENDIX D.

- o DATA SHOULD BE COLLECTED IN MEASUREMENT WINDOWS THAT ARE COMPARABLE FOR ALL THE WORKLOAD MEASURES USED IN THE CERTIFICATION EFFORT.

ANALYSIS RATIONALE

Once a valid, reliable, and applicable workload measure is used to assess flightcrew workload the process of reducing the data in order to understand and interpret the outcome begins. The appropriate statistical treatment of the workload data is necessary to yield interpretable results. The following section is not meant to be a treatise on statistical methods, rather it is intended to guide the sophisticated user towards the appropriate statistical method (for a useful review of statistical techniques see Kirk, 1982).

4.1 CANDIDATE SUBJECTIVE WORKLOAD MEASURES FOR AIRCRAFT CERTIFICATION

- The Bedford Scale (Modified Cooper-Harper type), the Subjective Workload Assessment Technique (SWAT), and the NASA Task Load Index (TLX) have demonstrated evidence of validity, reliability, and applicability as measures for evaluating pilot subjective workload post-flight.

- Although not evaluated in the simulation studies at NASA-Ames, comparative subjective evaluation techniques (e.g., Pilot Subjective Evaluation-PSE) have previously demonstrated applicability for

directly comparing two different aircraft, such as a baseline and a new aircraft.

As mentioned earlier, no workload measure can assess pilot workload in an absolute sense. Subjective measures have the same limit on interpretation in an absolute sense, but pilots tend to give the ratings as if there were an absolute scale underlying the rating. Pilots do not need a comparison to a baseline, such as a baseline aircraft, in order to generate workload ratings, but it is a rating of workload relative to a baseline of their earlier piloting experience. In this sense the subjective rating, a relative measurement instrument, is used as if it were an absolute measurement.

It is recommended that subjective workload assessment accompany a new aircraft certification effort because of the ability to query the user about the workload experienced during flight. The use of subjective measures has practical advantages (e.g., ease of implementation and non-intrusiveness). It should not be ignored that subjective measures of workload impinge additional workload on the pilot when a rating is requested. Biases in the rating process can also be introduced. Vested interest in aircraft certification, not wishing to appear as unable to handle a piloting situation, compliance with a biased probe (i.e., "this next flight segment is a no brainer,") are all situations which can bias an individuals subjective workload ratings.

4.1.1 IN-FLIGHT SUBJECTIVE MEASURE IMPLEMENTATION GUIDELINES

- PILOTS USED FOR THE WORKLOAD ASSESSMENT SHOULD BE UNAWARE OF THE MANIPULATION OF TASK DEMANDS (MALFUNCTIONS, DIVERSIONS, ETC.) DURING THE EVALUATION FLIGHTS.

- WHEN USING A SUBJECTIVE MEASURE IN-FLIGHT, THE MEASURE SHOULD NOT BE INTRUSIVE TO THE FLIGHT RELATED TASKS THE CREW MEMBER IS TRYING TO ACCOMPLISH.

- IF PAPER AND PENCIL RATING TECHNIQUES ARE TO BE USED IN FLIGHT, ONE CREW MEMBER AT A TIME SHOULD RECORD THEIR WORKLOAD RATINGS SO THAT THE OTHER CREW MEMBER MAY ATTEND TO FLIGHT DECK DUTIES.

- THE TO-BE-RATED FLIGHT SEGMENT (BEGINNING AND END POINTS) SHOULD BE CLEARLY IDENTIFIED TO THE FLIGHT CREW FOR THE PURPOSE OF OBTAINING THE DATA FOR EVALUATION.

Subjective ratings have been used to gather ratings of flightdeck activities in two different ways. One method asks pilots to consider a measurement period which is clearly defined over a period of time, the length of time is normally related to the task of interest. The other method is to request a subjective rating at a given instant in time. An instantaneous rating at that instant represents a snapshot of the workload encountered during the flight. Mechanical devices, boxes with labelled push button switches, have been employed in aircraft certification efforts to obtain instantaneous workload ratings (Speyer et. al., 1987; Wainwright, 1987).

These boxes normally have a cue light which illuminates when a rating is being requested, this method yields a snapshot or momentary response. These boxes could be modified so that a period is demarcated for evaluation, as opposed to a momentary rating. Traditional paper and pencil subjective rating techniques can be used in-flight, if the effect of intrusiveness is minimized.

4.1.2 POST-FLIGHT SUBJECTIVE MEASURE IMPLEMENTATION GUIDELINES

- o TO ENHANCE POST-FLIGHT WORKLOAD EVALUATION, VIDEOTAPE SHOULD BE USED TO AID THE CREW IN RECALLING THEIR SUBJECTIVE EVALUATIONS OF CREW WORKLOAD.

- o THE TO-BE-RATED FLIGHT SEGMENT (BEGINNING AND END POINTS) SHOULD BE CLEARLY IDENTIFIED TO THE FLIGHT CREW FOR THE PURPOSE OF OBTAINING THE DATA FOR EVALUATION.

- o WHEN USED, POST FLIGHT SUBJECTIVE RATINGS SHOULD BE COLLECTED FROM THE PILOTS AS SOON AFTER THE TASK AS OPERATIONALLY FEASIBLE.

A decision tree scale (e.g., Bedford, Modified Cooper-Harper, etc.) guides the rater in the application of the workload rating. The first decision point asks "Was it possible to complete the task?", if the answer is NO the rater is guided to the "10" rating. If the rater answers YES to the first decision point the rater will continue to the next decision point, "Was workload tolerable for the task?", if the answer is NO the rater is guided to the "7" through "9" ratings. If the rater answers YES to the second decision point the rater continues to the third decision point and are asked "Was workload satisfactory without reduction?", if the answer is NO the rater is guided to the "4" through "6" ratings. If the answer is YES to the last decision point the rater is guided to the "1" through "3" ratings.

The Subjective Workload Assessment Technique (SWAT) and the NASA Task Load Index (TLX) both use easy to implement computer-based mathematical methods for personalizing the assessment technique to reflect the idiosyncrasies of the rater. SWAT uses a card sort of the verbal descriptions of the 27 different ratings possible. The card sort activity can take anywhere from 20 to 60 minutes to complete. NASA-TLX uses a paired comparison selection technique, of the six bipolar scales, and requires approximately 5 minutes.

The PSE asks the rater to compare the new to be certificated aircraft with the aircraft on which he is currently rated (and flying) and evaluate the difference between the two aircraft on various systems (e.g., navigation, flight path control, aircraft systems, etc.). The systems called out by the comparisons correspond to the functions of FAR 25.1523 Appendix D.

4.2 CANDIDATE PHYSIOLOGICAL WORKLOAD MEASURES FOR AIRCRAFT CERTIFICATION

- Heart rate (in beats per minute or Inter-Beat Interval) has demonstrated evidence of reliability as a measure of pilot workload.

The use of physiological measures in certification is optimally accomplished by utilizing a direct comparison of baseline and to-be-certificated aircraft. The comparison can be accomplished in either simulation or flight test. An identical flight scenario should be used in order to eliminate any external factors influencing the workload comparison. The two aircraft should be flown by the same individuals to collect the physiological data. It is recognized that physiological measures, used in conjunction with subjective measures, can be used to identify changes in workload without a comparison to a baseline. As was mentioned before, no absolute measure of workload exists. Establishing criteria for an overload condition is a difficult task at best, but heart rate can be used to evaluate workload in a relative fashion.

Physiological measures must be collected on the same pilot for the baseline and new aircraft. Physiological measures are particularly sensitive to the differences between individuals. If the workload analysis was conducted on two different individuals on two different aircraft it would be impossible to tease apart differences in workload because of a change in pilots or a change in aircraft.

4.2.1 PHYSIOLOGICAL IMPLEMENTATION GUIDELINES

- o DATA COLLECTED WITH PHYSIOLOGICAL MEASURES CAN BE CONTAMINATED BY PHYSICAL MOVEMENT. SOURCES OF ARTIFACT SHOULD BE CONTROLLED WHEN EVALUATING THE IMPLEMENTATION OF A WORKLOAD MEASURE.

- o THE DATA SHOULD BE REPRESENTATIVE OF THE ENTIRE FLIGHT SEGMENT BEING EVALUATED, SO SOME SORT OF AVERAGING SHOULD BE USED WITHIN THE FLIGHT SEGMENT.

- o CARE SHOULD BE TAKEN SO THAT THE FLIGHT CREW IS PROTECTED FROM HAZARDS, SUCH AS ELECTRICAL SHOCK.

- o CARE SHOULD BE TAKEN TO ASSURE THAT THE PHYSIOLOGICAL ASSESSMENT METHOD APPEARS NON-CAREER THREATENING TO THE CREW MEMBERS IT EVALUATES (E.G., DATA COLLECTED USING PHYSIOLOGICAL MEASURES SHOULD CONTAIN NO DIAGNOSTIC MEDICAL INFORMATION).

Physiological measures can be obtained, stored, and statistically analyzed in a number of ways. However, equipment suited for medical purposes is not well suited for use in-flight or in a simulator. It is typically bulky, intrusive, requires the pilot to remain motionless, and requires 110 volt AC power. Appropriate specialized equipment is available from research supply firms. Consideration must be given to the accuracy of this equipment and the nature of the data collected. If analogue heart wave forms are collected it will be necessary to reduce, or digitize, the data for meaningful analysis. If digitized physiological

measures are rendered by specialized research equipment it is important to consider the time frame for which the data is collected (i.e., every 5 seconds, every 10 seconds, etc.).

Heart rate will increase, or Inter-Beat Interval (IBI) will decrease, with an increase in workload. It is more intuitive to consider heart rate in terms of beats per minute (i.e., 72 BPM). Inter-Beat Interval (IBI), the distance between spikes in an electrocardiogram record, can be transformed into heart rate using the following formula:

(1000 / Mean IBI milliseconds) * 60 = Heart Rate (BPM *)

* (BPM) Beats Per Minute

Heart rate has been found to increase as workload increases, e.g., during flight phases such as Takeoff and Landing, or when system malfunctions occur. Heart rate is affected not only by changes in physical workload, but also by mental workload.

Previously Roscoe (1983) has used heart rate data in the process of new aircraft workload certification with the CAA. Both have found changes in heart rate activity concomitant with changes in task demands. Hart (1987) argues, however, that heart rate may be a measure of arousal as opposed to workload.

4.3 CANDIDATE PERFORMANCE (PRIMARY TASKS) WORKLOAD MEASURES FOR AIRCRAFT CERTIFICATION

- Control input activity (i.e., wheel, column, and pedal) has demonstrated evidence of validity, reliability, and applicability as performance measures for evaluating pilot workload.

As was mentioned for physiological measures, the use of performance measures in certification is best accomplished by utilizing a direct comparison of baseline and to-be-certificated aircraft. The comparison can be accomplished in either simulation or flight test. An identical flight scenario should be used in order to eliminate any external factors influencing the workload comparison. The two aircraft must be flown by the same individuals to collect the necessary data.

A number of parameters related to aircraft control are tied with crew workload. The actual control activity involved with flying the aircraft has shown to be sensitive to changes in workload. The activity involved with manual flight path control is a valid indicator of workload. As task demands draw attention away from the primary action of piloting the airplane, the correction required to bring the airplane back onto the intended flight path becomes more pronounced. The more turns and altitude changes in a flight segment caused by these corrections using manual flight path control, the higher the workload.

In test aircraft most flight controls and flight surface position information is recorded during certification flights. One way in which performance can be operationalized is to examine significant control input activity over time.

4.3.1 PRIMARY TASK IMPLEMENTATION GUIDELINES

- o CONTROL INPUT ACTIVITY SHOULD BE EVALUATED ONLY DURING MANUAL FLIGHT PATH CONTROL.

- o WHEN POSSIBLE, STATE VARIABLES (E.G., PITCH ANGLE, ROLL ANGLE, ALTITUDE) SHOULD BE RECORDED CONTINUOUSLY IN SIMULATION TESTS.

- o WHEN POSSIBLE, WHEEL (AILERON) AND STICK (ELEVATOR) INPUTS SHOULD BE EMPLOYED TO REPRESENT AIRCRAFT CONTROL WORKLOAD THROUGHOUT THE ENTIRE FLIGHT OF AN AIRCRAFT UNDER MANUAL FLIGHTPATH CONTROL.

- o PEDAL (RUDDER) ACTIVITY IS NORMALLY ONLY REPRESENTATIVE OF AIRCRAFT CONTROL IN THE TAKEOFF AND APPROACH/LANDING PHASE OF THE FLIGHT AND SHOULD BE COLLECTED DURING THESE FLIGHT PHASES.

- o THE SAME FLIGHT SCENARIO SHOULD BE USED WHEN COMPARING NEW AND BASELINE AIRCRAFT.

- o A FLIGHT SHOULD BE DIVIDED INTO SEGMENTS FOR DATA COLLECTION SO DESCRIPTIVE STATISTICS (E.G., WHEEL POSITION, COLUMN POSITION) CAN BE COMPUTED ON THE CONTINUOUS MEASURES WITHIN EACH SEGMENT.

In addition to control activity, as previously defined operationally, control reversals of the wheel and control column are highly reliable measures for evaluation of pilot workload. The appropriate algorithm for computing control reversals is to follow the direction of activity in the flight controls, and when a change is made from one direction to another (i.e., column aft--climb, to column forward--descent), the counter should be incremented, and divided by units of time. The hysteresis, play in the controls (i.e., deadband), should be controlled for in the evaluation of both control inputs and reversals.

The position information needs to be collected at a fast enough rate to reflect real time control activity. A reasonable rate of collection of position information would be at a 10-hertz rate (10 updates per second). The control input activity is well represented by an algorithm that increments a counter by a movement in the control position of at least 2.5% every 10 hz has been found to be successful (Volume 1). The total amount of inputs needs to be divided by some unit of time (i.e., minutes) to yield a workload measure that can compare assessment periods of varying durations.

4.3.2 SECONDARY TASKS

Secondary tasks have proven to be reliable, valid, and applicable in laboratory studies of workload. Many techniques exist (e.g., time estimation, mental mathematics, choice-reaction time, critical tracking task, and Sternberg tasks). Secondary tasks are sensitive to the spare capacity available for performing tasks during workload evaluation. As the resources available are being drawn upon for the completion of tasks, the spare capacity for additional tasks diminishes.

4.3.2.1 SECONDARY TASK IMPLEMENTATION GUIDELINES

- o WHEN USED, SECONDARY TASKS SHOULD BE EMBEDDED IN THE FLIGHT TASK SO AS TO BE AS NON-INTRUSIVE AS POSSIBLE.

- o EMBEDDED SECONDARY TASKS SHOULD NOT APPEAR ARTIFICIAL TO THE OPERATOR SO AS TO MAINTAIN OPERATOR ACCEPTANCE AND FACE VALIDITY.

- o SECONDARY TASKS ARE MOST EFFECTIVELY IMPLEMENTED IN A SIMULATION ENVIRONMENT, WHERE AIR SAFETY IS NOT A CONCERN AND CONTROL OF THE ENVIRONMENT IS POSSIBLE.

- o SECONDARY TASK TECHNIQUES SHOULD BE AVOIDED WHEN INTRUSION WILL SERVE AS A SOURCE OF INTERFERENCE FOR THE PRIMARY WORKLOAD MEASURES.

A benefit to using secondary tasks as a measure of workload is the capability of detecting variations in non-overload conditions. Performance on the secondary task does vary with increasing levels of workload, unlike traditional primary measures which tend to degrade abruptly when overload occurs. An additional benefit is the means provided for diagnosing what actions are causing increases in workload:

(a) Hart (1978), Gunning (1978), Wierwille and Connor (1983), and Casali and Wierwille (1983) have used time estimation as a secondary task method to assess workload in aviation related environments.

(b) Huddleston and Wilson (1971), Green and Flux (1976), and Wierwille and Connor (1983) have used mental mathematics as a secondary task method to assess workload in aviation related environments.

(c) Kantowitz et. al. (1983, 1984) have used choice reaction time as a secondary task method to assess workload in aviation related environments.

(d) Jex and Clement (1979), and Burke et. al. (1980) have used critical tracking tasks as a secondary task method to assess workload in aviation related environments.

(e) In the past many researchers have used the Sternberg task to evaluate workload in the aviation environment. O'Donnell (1976), Crawford et. al. (1978), Wolfe (1978), Wickens and Derrick (1981), Schifflet et. al. (1982), and Wierwille and Connor (1983) all have employed the Sternberg task in evaluating workload in aviation related paradigms.

(f) Shingledecker and Crabtree (1982), and Silverstein, et. al. (1984) have used embedded radio probes as a methodology in aviation related environments. In addition to ATC probes being used to

implement the secondary task, squawking and identing new transponder codes has been used successfully in simulation environments.

4.4 CANDIDATE ANALYTICAL ASSESSMENT TECHNIQUES FOR AIRCRAFT CERTIFICATION

- The analytical assessment technique of Timeline Analysis has demonstrated evidence of validity and applicability for assessing crew task demands.

Analytic assessment techniques (e.g., task timeline analyses) offer an effective tool for estimating crew task-demands on competing design alternatives. In addition to an aid in design, timeline data has been used effectively in evaluating flightdeck workload for aircraft certification.

Analytical techniques are of particular value to the aircraft manufacturer since they offer both the potential for identifying and correcting workload problems early in the design phase when the cost of change is relatively low and a tool which can provide data for certification. Design of a complex flight-deck is an iterative process. Early in the development process, many details about its functioning need to be specified.

After a conceptual design phase has been completed and some basic decisions have been made about the functional allocation of tasks between man and automation, a detailed description of the required crew functions is needed before the designer can outline the crew interface (e.g., controls and displays). Functions of system monitoring, assessment, decision making, and operation of controls is determined at the greatest level of detail that is practical and accurate. Obviously, greater levels of accurate detail will yield higher levels of confidence that the final flight deck design will have acceptable levels of crew task-demands, and consequently, acceptable levels of crew workload.

Many variations of the task analysis technique have been developed because it is an adaptable and cost-effective tool for design and evaluation (Parks, 1978; Stone, Gulick, and Gabriel, 1985). Timeline analyses enable predictions to be made regarding the likely workload of a new system.

The relative (as opposed to absolute) nature of workload measures has led airplane manufacturers to demonstrate workload acceptability during certification by means of a relative comparison between new and existing aircraft models. The task-demands of the new model aircraft are generally compared to levels found in existing aircraft which have a similar flight profile (task-demands) and a good safety record (Fadden, 1982; Stone, Gulick, and Gabriel, 1985). Conclusions regarding workload acceptability are generally based on relative workload comparisons showing the same or reduced workload (or task demand) levels on the new aircraft when compared to the model which was already in service. When performing a relative comparison with a new aircraft design, however, there may be instances when workload levels exceed the old design. In cases such as this, the increased workload is not necessarily unacceptable, but it may become the subject of a more in-depth workload analysis. These cases need to be considered on a case-by-case basis with all of the operational factors taken into consideration when evaluating the impact of small workload increases. In some

cases, small increases in crew workload may actually be desirable. When using the task timeline analysis technique, the task demands of a new model aircraft are compared to levels found in existing aircraft that have a similar flight profile (task demands) and a good safety record.

4.4.1 ANALYTIC TECHNIQUES IMPLEMENTATION GUIDELINES

- WHEN USING ANALYTIC TECHNIQUES, CONCLUSIONS REGARDING WORKLOAD ACCEPTABILITY SHOULD BE BASED ON RELATIVE WORKLOAD COMPARISONS BETWEEN THE TO-BE-CERTIFICATED AIRCRAFT AND BASELINE.

- WHEN PERFORMING A RELATIVE COMPARISON WITH A NEW AIRCRAFT DESIGN AND WORKLOAD LEVELS EXCEED THE OLD DESIGN FOR A FLIGHT SEGMENT, A DECISION SHOULD BE MADE REGARDING THE NEED FOR A MORE IN-DEPTH WORKLOAD ASSESSMENT.

- DETAILED PROCEDURES SHOULD BE DEVELOPED THAT DEFINE ALL ACTIONS EACH CREW MEMBER MUST ACCOMPLISH TO COMPLETE A FLIGHT SUCCESSFULLY.

- CREATION OF THE SCENARIO SHOULD BE BASED UPON DATA DERIVED FROM FLIGHT PLANS, NAVIGATION CHARTS (SID, STAR, ENROUTE AREA, APPROACH, AND AIRPORT MAPS), ATC OPERATIONAL DATA, AIRCRAFT PERFORMANCE DATA, AND AIRCRAFT OPERATIONS MANUALS.

Timeline analysis depends upon a time based flight scenario which describes the discrete crew procedures associated with flying a route. The set of flight procedures to be accomplished and the time available for performance are directed by the scenario. Creation of the scenario is based upon data derived from flight plans, navigation charts, (SID, STAR, enroute, area, approach and airport maps), ATC operational data, aircraft performance data, and aircraft operations manuals. Using this information base, detailed procedures are developed which define all actions each crew member must accomplish to successfully complete a flight. Task timeline analyses enable the designer to make relatively conservative estimates of what the crew's task-demands will be so that it will be virtually certain that the actual workload experienced by the crew will be acceptable.

The time descriptions employed in task analyses techniques (Miller, 1976; Stone, Gulick, and Gabriel, 1985) have been extensively refined over the years to improve their accuracy to enable more accurate detail designs. In many cases the task-times are based on measurements of actual crew performance so that the task analyses will provide the best possible estimate of actual crew behavior in the finished airplane. In fact, it was concluded by the President's Task Force on Crew Complement that the timeline analyses performed by airframe manufacturers represented the state-of-the-art during the last transport airplane certification efforts (McLucas, Drinkwater, and Leaf, 1981). Because the task times in the task timeline analyses were validated with actual crew performance data, it was concluded they were representative of the actual crew workload that would be experienced by any trained flight crew. It is worth emphasizing that

timeline analysis provides invaluable estimates of crew task-demands during design and provides a useful means of supporting workload certification.

5.0 TASK SCENARIO DEVELOPMENT

When the certification of an aircraft occurs, flightdeck workload should be evaluated under operating conditions which are as realistic as possible during simulation or flight test. It is during high fidelity simulation or actual operation that workload measures can confirm predictions made by analytic methods, such as task timeline analysis, that estimated workload levels are reasonable. It is imperative, therefore, that the task scenarios developed for certification activities are representative of airline operations.

5.1 SCENARIO DESCRIPTION GUIDELINES

- o DISCRETE MEASUREMENT PERIODS SHOULD BE USED FOR EVALUATING WORKLOAD, OTHERWISE SPECIFIC EVENTS OR ACTIONS ARE MIXED WITH UNWANTED TYPES OF WORKLOAD IRRELEVANT TO THE CONCERNS FOR CERTIFICATION. THIS LEADS TO THE INABILITY TO EVALUATE DISCRETE VARIATIONS IN CREW WORKLOAD.

The level of detail of the scenario description should allow discussion of all the intended actions that are to be evaluated with the to-be-certificated aircraft. Detailed procedures should be developed that define all actions each crew member must accomplish to complete a flight successfully.

5.2 SCENARIO EVALUATION CRITERIA

5.2.1 ROUTE CHOICE

- o THE ROUTES SHOULD PROVIDE A REPRESENTATIVE MIX OF NAVIGATION AIDS, AIRPORTS, APPROACHES, AND AIR TRAFFIC CONTROL (ATC) SERVICES. IN ADDITION, ROUTES THAT ADEQUATELY SAMPLE HIGH DENSITY AREAS.

5.2.2 NEW TECHNOLOGY AND EQUIPMENT

- o THE SCENARIO FLOWN SHOULD TAKE INTO ACCOUNT THE NEW EQUIPMENT INCORPORATED INTO THE TO-BE-CERTIFICATED AIRCRAFT. PROCEDURES ASSOCIATED WITH THE NEW EQUIPMENT SHOULD BE ADDRESSED, AS WELL AS OPERATIONAL AND MINIMUM EQUIPMENT LISTS.

5.2.3 CRITICAL EVENT IDENTIFICATION

- o THE SCENARIO FLOWN SHOULD REPRESENT THE RANGE OF OPERATIONAL REQUIREMENTS OF THE TO-BE-CERTIFICATED AIRCRAFT INCLUDING REPRESENTATIVE NORMAL AND NON-NORMAL PROCEDURES LIKELY TO BE ENCOUNTERED DURING ITS OPERATION IN SERVICE.

5.3 PROCEDURES USED TO RELATE WORKLOAD TO FAR REQUIREMENTS

- o WORKLOAD FUNCTIONS AND WORKLOAD FACTORS DESCRIBED IN FAR 25 APPENDIX D SHOULD BE REPRESENTED IN THE SCENARIOS FLOWN IN THE CERTIFICATION EFFORT.

Since the certification of an aircraft occurs under operating conditions which are as realistic as possible, any serious concerns about workload acceptability of a new aircraft can be evaluated during the flight test phase of certification. It is during high fidelity simulation or actual operation that workload measures provide information about representative crew performance to confirm that they can reliably cope when in service. In order to show compliance with FAR 25.1523, Appendix D, the scenario proposed should incorporate the following factors:

Creation of the scenario should be based upon data derived from flight plans, navigation charts (SID, STAR, enroute area, approach, and airport maps), ATC operational data, aircraft performance data, and aircraft operations manuals.

A sampling of representative non-normal procedures should be established in the test scenario to show their effect on the crew workload. The acceptability of all procedures should be verified, as well as the distribution of crew workload during the execution of these procedures. Critical items and reasonable combinations of inoperative items should be considered in dispatching the aircraft. In addition, the scenario should incorporate adverse weather such as that which the aircraft is likely to encounter during its operation in service.

Operational definitions of crew workload required for scientific testing and flight deck certification address two primary concerns on the part of the flight crew: (1) Is there sufficient time for the crew to accomplish all of the tasks required for operating the aircraft under the demands of line operational flight? and (2) Can this be done without causing undue mental or physical stress? The workload factors, which are identified in Appendix D, are described in general terms but they are not operationally defined.

Clear measurement of the workload factors has not been a simple or scientifically precise matter. The operational definitions of workload need to be clarified as they apply to the applicable workload factors mentioned in Appendix D.

The FAA has provided a list of what they consider to be the important types of workload for aircraft certification in FAR 25.1523, Appendix D (see TABLE 1.2-1). Design of the scenario for workload assessment should include the types of workload described therein. In fact, the test scenario should address the basic workload functions and factors listed in FAR 25.1523, Appendix D. For example, in an evaluation of communications workload, the scenario should include the basic communications required to properly operate the airplane in the environment for which approval is sought. Appropriate company and cabin communications should also be incorporated. The goal is to evaluate the specific types of workload with the appropriate crew complement during realistic operating conditions, including representative weather, air traffic, and airline operational duties.

The following summarizes the task of operationally defining the functions and factors of Appendix D: (1) identify and detail the mission segment task-demands

which pose a concern for transport aircrew workload, (2) categorize the operationally relevant types of task-demands which would be expected to occur as a result of the selected mission segments and mission profiles,a and (3) use a categorization scheme which employs the descriptions found in FAR 25.1523, Appendix D (see Table 1.2-1). An example of this procedure can be found in section 5.3 of Volume One of this Final Report.

6.0 DISCUSSION

Guidelines for evaluating an aircraft workload certification plan have been presented.

In addition, criteria for evaluating whether or not a workload assessment technique is valid and reliable is presented. If a proposed workload measurement technique does not have an empirical record of validity and reliability the process by which an applicant can prove justification for usage of the measure is presented.

Volume One (Assessment of Crew Workload Measurement Methods, Techniques, and Procedures. Volume One: Final Report) contains the results of simulation studies conducted at NASA-Ames. The simulation studies reported includes the first attempt at collecting test-retest reliability data for workload measures in a full fidelity simulation (ATC included) of the commercial transport environment.

Further work must be conducted to allow for more specific measurement of mental workload. Aircraft flightdecks are becoming more automated, which requires the flightcrew to take on the role of system monitors. Additional work must be conducted to eliminate the effect of individual differences in order to allow for the determination of the point at which overload occurs across the pilot population.

Independent of the issues to be investigated in further workload studies, it is hoped that the precedence established for evaluating the validity and reliability of assessment techniques will be continued.

REFERENCES

Anastasi, A. (1968), Psychological Testing (3rd ed.) London: Macmillan.

Burke, M. W., Gilson, R. D., and Jagacinski, R. J. (1980), "Multimodal Information Processing for Visual Workload Relief," Ergonomics, 23, 961-975.

Casali, J. G., and Wierwille, W. W. (1983), "Communications-Imposed Pilot Workload: A comparison of Sixteen Estimation Techniques," Proceedings of Second Ohio State University Symposium on Aviation Psychology, 223-235.

Childress, M. E., Hart, S. G., and Bortolussi, M. R. (1982), "The Reliability and Validity of Flight Task Workload Ratings," Proceedings of the Twenty-sixth Annual Meeting of the Human Factors Society.

Crawford, B. M., Pearson, W. H., and Hoffman, M. (1978), "Multipurpose Digital Switching and Flight Control Workload," Report No. AMRL-TR-78-43. Wright-Patterson Air Force Base, Ohio: USAF Aerospace Medical Research Laboratory.

Damos, D. (1984), "Classification Systems for Individual Differences in Multiple-Task Performance and Subjective Estimates of Workload," Proceedings of the Twentieth Annual Conference on Manual Control, Sunnyvale, CA.

Eggemeier, F. T., Crabtree, M. S., and Reid, G. B. (1982), "Subjective Workload Assessment in a Memory Update Task," Proceedings of Twenty-sixth Annual Meeting of the Human Factors Society.

Fadden, D. M. (1982), "Boeing Model 767 Flight Deck Workload Assessment Methodology," Paper presented at the SAE Guidance and Control System Meeting, Williamsburg, VA.

Green, R., and Flux, R. (1976), "Auditory Communication and Workload," Proceedings of the AGARD Conference on Methods to Assess Workload, (AGARD-CPP-216), A4/1-A4/8.

Gunning, D. (1978), "Time Estimation as a Technique to Measure Workload," Proceedings of the Twenty-Second Annual Meeting of the Human Factors Society, 41-45.

Hart, S. G. (1978), "Subjective time estimations an index of workload. Proceedings of Airline Pilots Association Symposium on Man-System Interface: Advances in Workload Study, 115-131, Washington, D.C..

Hart S. G., and Bortolussi, M. R. (1983), "Pilot Errors as a Source of Workload," Proceedings of the Second Symposium on Aviation Psychology, Columbus, OH.

Hart, S. G. (1987), "Measurement of Pilot Workload," Proceedings of the AGARD Conference on Methods to Assess Workload, (AGARD-CPP-282), 116-123.

Huddleston, H. F., and Wilson, R. V. (1971), "An Evaluation of the Usefulness of Four Secondary Tasks in Assessing the Effect of a Lag in Simulated Aircraft Dynamics," Ergonomics, 14(3), 371-380.

Jex, H. R., and Clement, W. F. (1979), "Defining and Measuring Perceptual-Motor Workload in Manual Control Tasks," In N. Moray (Ed.), Mental Workload: Its Theory and Measurement. New York: Plenum.

Kantowitz, B. H., Hart, S. G., Bortolussi, M. R., Shively, R. J., and Kantowitz, S. C. (1984), "Measuring Pilot Workload in a Moving-Base Simulator: Building Levels of Load," Proceedings of the 20th Annual Conference on Manual Control, (NASA CP-2341), 359-372.

Kantowitz, B. H., Hart, S. G., and Bortolussi, M. R. (1983), "Measuring Pilot Workload in a Moving-Base Simulator: Asynchronous Secondary Choice-Reaction Task," Proceedings of the Twenty-seventh Annual Meeting of the Human Factors Society, 319-322.

Kirk, R. E. (1982), Experimental Design: Procedures for the Behavioral Sciences. Monterey: Brooks/Cole.

McLucas, J. L., Drinkwater, F. J., and Leaf, H. W. (1981), "Report of the President's Task Force on Aircraft Crew Complement," Prepared for the President, The White House, Washington D. C..

Miller, K. M. (1976), "Timeline Analysis Program (TLA-1), final report," Boeing Document D6-42377-5, Prepared for National Aeronautics and Space Administration, Langley Research Center (NASA-CR-144942).

O'Donnell, R.D. (1976), "Secondary Task Assessment of Cognitive Workload in Alternative Cockpit Configurations," In B.O. Hartman (Ed.), Higher mental functioning in operational environments (AGARD Conference Proceedings No. 181). Nevilly sur Seine, France: Advisory Group for Aerospace Research and Development.

Parks, D. L. (1978), "Current Workload Methods and Emerging Challenges," In N. Moray (Ed.), Mental Workload: Its Theory and Measurement. New York: Plenum Press.

Roscoe, A.H. (1983), "Analysis of Heart-Rate Data," In: Certification Report British Aerospace, HTD R 460-00 SC0038 Annex K.

Schifflet, S.G., Linton, P.M., and Spicuzza, R.J. (1982), "Evaluation of a Pilot Workload Assessment Device to Test Alternative Display Formats and Control Handling Qualities," Proceedings of the 1982 AIAA Workshops on Flight Testing to Identify Pilot Workload and Pilot Dynamics, 222-233.

Shingledecker, C. A., and Crabtree, M. S. (1982), "Standardized Tests for the Evaluation and Classification of Workload Metrics," Proceedings of the Twenty-sixth Annual Meeting of the Human Factors Society.

Silverstein, L.D., Gomer, F.E., Crabtree, M.S., and Acton, W.H. (1984), "A Comparison of Analytic and Subjective Techniques for Estimating Communications-Related Workload During Commercial Transport

Operations," (NASA CR-2341), Washington, D.C.: National Aeronautics and Space Administration.

Speyer, J. J., Fort, A., Fouillot, J. P., and Blomberg, R. D. (1987), "Assessing workload for minimum crew certification," Proceedings of the AGARD Conference on Methods to Assess Workload, (AGARD-CPP-282), 90-115.

Stone, G., Gulick, R. K., and Gabriel, R. F. (1985), "Use of Task/Timeline Analysis to Assess Crew Workload," Douglas Paper 7592. Douglas Aircraft Company, Long Beach, CA.

Wainwright, W. A. (1987), "Flight Test Evaluation of Crew Workload," Proceedings of the AGARD Conference on Methods to Assess Workload, (AGARD-CPP-282), 60-69.

Wickens, C.D., and Derrick, W. (1981), "Workload Measurement and Multiple Resources," Proceedings International Conference on Cybernetics, 600-603.

Wierwille, W. W., and Casali, J. G. (1983), "A Validated Rating Scale for Global Mental Workload Measurement Applications," Proceedings of the Twenty-seventh Annual Meeting of the Human Factors Society.

Wierwille, W.W., and Connor, S.A. (1983), "Evaluation of 20 Workload Measures using a Psychomotor Task in a Moving-Base Aircraft Simulator. Human Factors, 25, 1-16.

Wolfe, J.D. (1978), "Crew Workload Assessment: Development of a Measure of Operator Workload," Report No. AFDL-TR-78-165. Wright-Patterson Air Force Base, Ohio: Air Force Flight Dynamics Laboratory.

☆U.S. GOVERNMENT PRINTING OFFICE: 1989 - 748-056/24038